An Algorithm for correcting CoRoT raw light curves

D. Mislis[1], J.H.M.M. Schmitt[1], L. Carone[2], E. W. Guenther[3] & M. Pätzold[2]

[1]Hamburger Sternwarte, Gojenbergsweg 112, D-21029 Hamburg, Germany

[2] Rheinisches Institut für Umweltforschung, Abteilung Planetenforschung, an der Universität Köln, Aachener Str. 209, 50931 Köln, Germany

[3]Thüringer Landessternwarte Tautenburg, Sternwarte 5, D-07778 Tautenburg, Germany

e-mail: mdimitri@hs.uni-hamburg.de

ABSTRACT

We introduce the CoRoT Detrend Algorithm (*CDA*) for detrending CoRoT stellar light curves. *CDA* has the capability to remove random jumps and systematic trends encountered in typical CoRoT data in a fully automatic fashion. Since huge jumps in flux can destroy the information content of a light curve, such an algorithm is essential. From a study of 1030 light curves in the CoRoT IRa01 field we developed three simple assumptions CDA is based upon. In this paper we describe analytically the algorithm and we provide some examples of how it works. We demonstrate its functionality of the algorithm in the cases of CoRoT0102702789, CoRoT0102874481, CoRoT0102741994 and CoRoT0102729260. Using CDA in the specific case of CoRoT0102729260 we detect a candidate exoplanet around the host star of spectral type G5, which remains undetected in the raw light curve; the estimated planetary parameters are $R_p = 6.27R_E$ and $P = 1.6986$ days.

Key words. methods: data analysis, surveys, planetary systems, stars: variables

1. Introduction

The CoRoT satellite was successfully launched in 2006. On board CoRoT there is a small 27cm telescope feeding two science channels to study astroseismology and transits respectively (Baglin et al. 2000). The CoRoT has a field of view (FOV) of $\sim 2.7^o$ x 3.05^o. In its first field (IRa01 - $\alpha = 6^h46^m53^s$ & $\delta = -00^o12'00"$), CoRoT had observed continuously for 60 days, producing uninterrupted light curves for the first time. The data from the IRa01 have been public since December 2008 and the astronomical community has access to these data. Unfortunately, the CoRoT light curves are affected by a variety of instrumental problems, which severely hamper the data interpretation. In order to overcome these difficulties we have developed the CoRot Detrend Algorithm (*CDA*). In this paper the algorithm is presented and demonstrate its function on some typical CoRoT data sets.

2. CoRoT light curves: The problems

The CoRoT data files contain multi-color light curves, produced by inserting a low-resolution dispersing prism into the telescope beam. With this set-up it is intended to provide simultaneous light curves in the red (R), green (G) and blue (B) bands, however, these bands do not correspond to true photometric filters and, in fact, the bands may differ from star to star. We study the multi-color data in this paper, but also consider the total (white) flux obtained by summing up the individual light curves through $W = R + G + B$.

Fig. 1 are typical CoRoT light curves from IRa01. The first panel of Fig. 1 shows a typical exponential jump very similar to a flare star. A trend is also evident. In the second light curve there appears a box-shape jump, in the third and fourth light curves one finds features similar as in the first and second light curves, except that the jumps are downwards. We note that the downward jump in the third light curve is very similar to a transit event, thus making the detection of true transits difficult. Combinations of all the above features appear, in fact a rather typical CoRoT light curve. Essentially, two basic instrumental problems appear in all CoRoT light curves: First, there is a long-term trend, forcing a secular decrease of the light curve intensity over the full observing period of 60 days. The strengths of the trends in different sources may be different; the physical cause of these trends is not well understood. The second and even more serious problem is the instrumental jumps in the light curves. The term "jump" refers to a sudden variation of intensity without any obvious reason. Many of these jumps do in fact look like stellar flares, however, the vast majority of these features is clearly instrumental. The physical explanation for these jumps could be, cosmic radiation and the time evolution of bright pixels (Pinheiro da Silva et al. 2008). These jumps are a random phenomenon and affect each filter differently. An inspection of hundreds of CoRoT light curves similar to those presented in Fig. 1 allows to classify the observed shapes of jumps into five groups:

- Sudden intensity increase and exponentially decrease (Fig. 1 - panel a)
- Sudden intensity increase and decreases (box shape, Fig. 1 - panel b)
- Sudden intensity decrease and exponentially increase afterwards (Fig. 1 - panel c)
- Sudden intensity decrease and increase (negative box shape, Fig. 1 - panel d)
- All of the combinations above (Fig. 1 - panel e)

A statistical analysis of IRa01 field (visual inspection) shows that only a small minority (Table 1) of all jumps is so powerful that they simultaneously appear in each colour. Most of the light
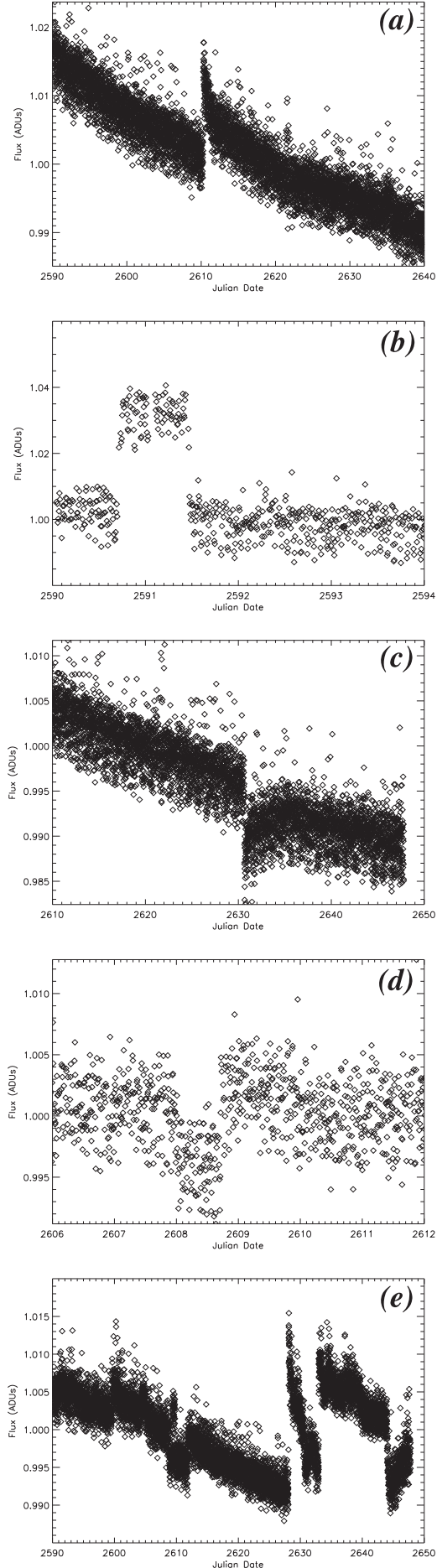
Table 1. Statistical analysis of 1030 CoRoT light curves from IRa01. Jumps appear in more than 50% of all light curves in all filters; in 0.82% of all light curves jumps in all filters occur at the same time.

R filter	G filter	B filter	Total
38.14%	14.4%	15.1%	67.6%

curves are affected not only by one single jump, but by many jumps occurring in the different filters at different times. In Table 1 we show the results of a statistical study of the appearance and the shapes of jumps using data form IRa01. The three first columns of Table 1 show the number of light curves which suffer from jumps in the respective filter filter and the fourth column shows the total amount.

3. The CDA Algorithm

3.1. General features

It is quite difficult to describe all the features perturbing CoRoT light curve with a given function, since there are many different shapes of jumps with many different functional forms. Furthermore, the problem is complex, because we do not know which of light curve features are real signals (real transits, real flares etc.) or instrumental effects. The algorithm is based on three assumptions: (a) trends appear in almost all light curves and both flux increases and decreases can occur. The trends are not periodic and we assume them to be a long-term phenomenon (Aigrain et al. 2009). (b) The second assumption also accrues from the statistical analysis of the data. The study of 1030 light curves from IRa01 field shows that only 0.82 % of them are affected by a jump in all three filters at the same time. In these cases the jump is very large and affects all bands with the same temporal pattern, however, in most cases the jumps affect only one band at any given time (Fig. 2), and we therefore ignore those cases where jumps occur simultaneously in all three bands. (c) Real transits must appear in all three filters, while, of course, the intensity and transit depth can vary from filter to filter. In summary, for the *CDA* we assume that

– Long term trends appear in all CoRoT light curves
– Jumps are random phenomena appearing in different filters at different times.
– The real signals from transits appear in all three bands

We emphasize that *CDA* works only for events (like transits), which appear in two or more bands; *CDA* does not work for stellar flares, since most stellar flares do not show any flux enhancements in the red and green band, but in the blue band. Under these circumstances *CDA* will destroy real signals, unless the flare is so powerful to appear in all bands.

3.2. The algorithm

CDA uses all the colour light curve simultaneously of each star to remove the instrumental features. The basic idea of *CDA* is to use the cleanest filter band as a proxy for the whole light curve. The raw data files of each CoRoT light curve have a quality flag (CoRoT files - column 4), indicating the quality of each data point (Mazeh et al. 2009). We first remove all these "bad points" (points with high noise flagged by CoRoT); note that these "bad points" are same for all the filters per star. In this paper we will
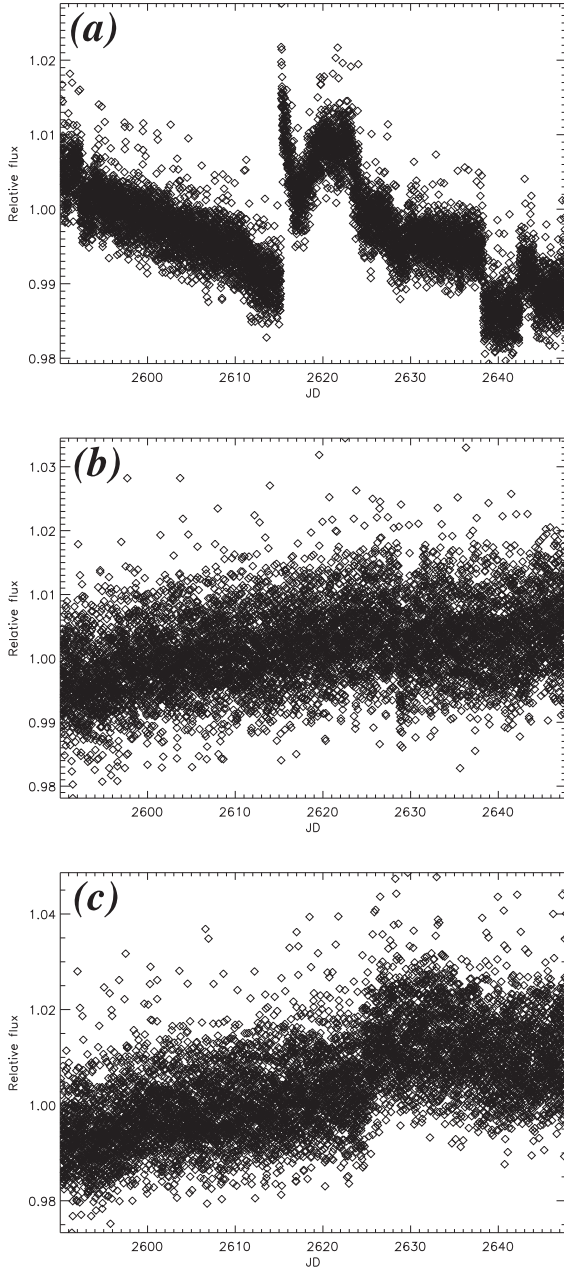
Fig. 1. Jumps and trends in CoRoT light curves. CoRoT01027-21492
(a) -24482 (b) -40879 (c) -49307 (d) - 27431 (e)

Fig. 3. Standard deviation vs number of blocks.

After this step *CDA* proceeds to remove the jumps. In order to identify the cleanest light curve for a reliable jump removal we create "sub-light curves", which we typical take with a duration of a day. Thus, for the IRa01 field we create 60 "sublight" curves, called simply light curves in the following. These 60 blocks were selected after we checked various combinations. If the number of blocks are too large, then transit signals are reduced, and if the number of blocks are too small, the probability to include a jump in the "sublight" curve increases. Fig 3 shows the best block number vs standard deviation.

Let us assume that there are three full light curves for a given star in each band with N points per light curve; denote by $F_{R,i}$, $F_{G,i}$ and $F_{B,i}$ with $i = 1, N$ the individual data values in the red, green and blue filters, respectively. Then we divide each color light curve in 60 sub-light curves (one sub-light curve per day for IRa01 - 60 days). For each sub-light curve we calculate the mean value *MR*, *MG* and *MB* and normalize each sub-light curve by its mean value; we compute new, normalized sub-light curves *NF* through

$$NF_{R,G,B,i} = \frac{F_{R,G,B,i}}{M_{R,G,B}} \qquad (2)$$

for each filter band and it is clear that all of these light curves have a mean of unity. This normalization is necessary since otherwise the whole process would be dominated by the light curve with the highest signal, which is usually the red light curve. As a side effect, *CDA* normalizes the depth of a possible transit in all filters using equation 2, so when the algorithm continues with its next steps, all transit events in each filter will have the same depth and thus *CDA* does not destroy real signals from the transits.

The normalized light curves have now the same mean, their dispersions will, however, differ. Our next goal is to identify the instrumental scatter, caused, for example by jumps, in each light curve and disentangle this instrumental scatter from statistical noise. In order to achieve this, *CDA* extracts five random packages of twenty adjacent points each from all colour bands and calculates the standard deviation of each package per filter; the result should represent a good estimate of the correct light curve value at that time. If we use many packages the probability to include jumps increases. The correct combination packages-points is a function of the duration of the jumps which is a random value, thus there is no a fix combination. We define as the mean

Fig. 2. CoRoT0102729260. Three filter light curves (R **(a)**, G **(b)**, B **(c)**) from a data set. The jumps in red light curve does not appear in the other filters and vice versa.

use light curves with all "bad points" already removed (as in Fig. 1). As noted above in our first assumption, trends are a long-term phenomenon. A 3rd degree polynomial is fit to the entire light curve in order to remove the trend in each filter per star. Because each CoRoT light curve typically has thousands of data points, the polynomial does not fit short-term variations and real short-term events like transits. We thus write

$$Flux = a + b \cdot JD + c \cdot JD^2 + d \cdot JD^3, \qquad (1)$$

where JD is the Julian date (normalized to range $-1 \leq JD \leq 1$) and a, b, c and d are the fit parameters for the third degree polynomial. At the end of this procedure, we have a detrended light curve per filter for each star.
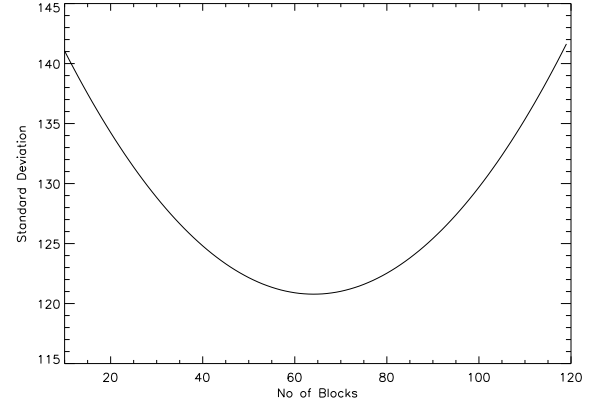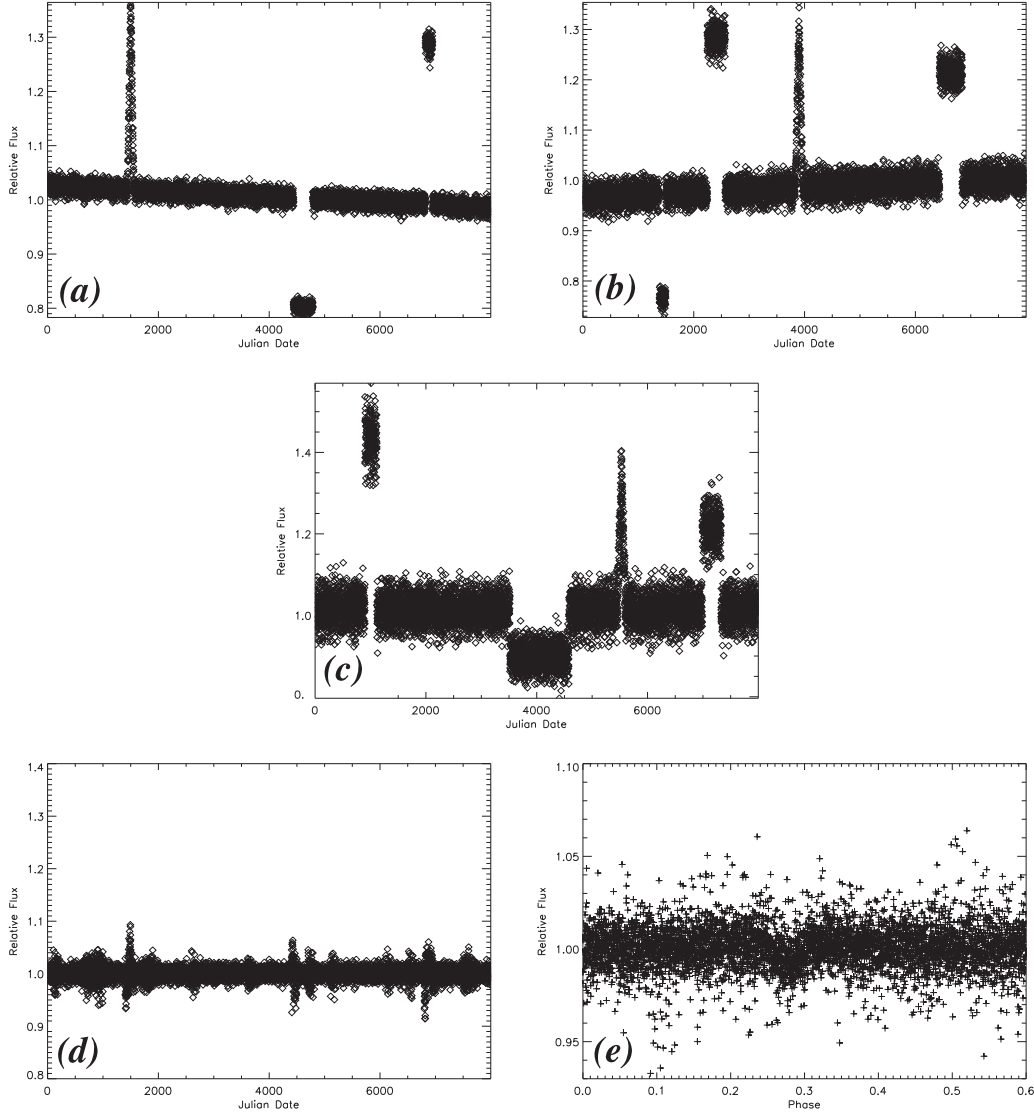
Fig. 4. Simulated data. R - **(a)**, G - **(b)** and B -**(c)** color respectively. Plot **(d)** is the final light curve after CDA and the plot **(e)** is the phase diagram of the transit after CDA & BLS.

standard deviation (MSD), the mean value of these five packages of each filter

$$MSD_{R,G,B} = \frac{1}{5} \sum_{j=1}^{5} \frac{1}{20} \sqrt{\sum_{i=k_j}^{k_j+20} (NF_{R,G,B,i} - Mean_{min})^2}, \quad (3)$$

where the induces k_j denotes 5 different random data points of the light curve and $Mean_{min}$ is the mean value of the flux of each package. In general, each filter has a different MSD value, which is compared with the standard deviation of each filter TSD defined through

$$TSD_{R,G,B} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (NF_{R,G,B,i} - Mean_{min})^2} \quad (4)$$

Finally, the relative standard deviation of each filter RSD is computed and defined by

$$RSD_{R,G,B} = \frac{TSD_{R,G,B}}{MSD_{R,G,B}}. \quad (5)$$

At the end of this process we have three normalized light curves $NF_{R,i}$ $NF_{G,i}$ and $NF_{B,i}$, and three values for the relative standard deviation RSD_R, RSD_G and RSD_B for each filter light curve respectively. *CDA* compares these three numbers and calls the light curve with the minimum RSD the base and the light curve with the maximum RSD, target. To make the procedure more understandable we continue with an example: Suppose the base is the blue light curve ($NF_{B,i}$) and the target is the red ($NF_{R,i}$) light curve. Using base and target *CDA* calculates a new mean light curve (AF_i); in our example *CDA* computes

$$AF_i = \frac{1}{2}(NF_{R,i} + NF_{B,i}). \quad (6)$$

and then it recalls the AF_i as the light curve with the maximum RSD (in this example recall AF_i as $NF_{R,i}$). According to assumptions 2 and 3, in the AF_i light curve remains any possible real signal but all the fake (jumps) tend to be reduced, because jumps appear only at specific times in each filter. As a final result we will have a red light curve reduced and two others (green and blue) untouched. If we try to run the algorithm again we will notice that the new values of RSD have changed because one light

curve has changed. This means that every time we run the previous step of the algorithm, *CDA* removes a part of a fake signal (Fig. 3).

When these loops end, we re-normalize the final light curve of the red channel to the raw mean value,

$$NFR_{final} = NR \cdot NF_{R,i} \qquad (7)$$

and the procedure has been completed. NFR_{final} is the final sub-light curve. The final step is to put all the 60 sub-light curves together. This is the final light curve and we are ready to search for exoplanets (Fig. 5). Of course we use many loops for procedure, but if we use too many, *CDA* starts to destroy the light curve because it is obvious that after some loops there is a "saturation" in the procedure. To avoid this effect, we do not use the same loop number of each light curve. We calculate the standard deviation of each light curve after each loop and *CDA* stops when the standard deviation starts to increases.

3.3. Simulations

In order to verify the functionality of CDA, we simulated CoRoT light curves as shown in Fig. 4. We specifically simulated a light curve in three filters (R,G,B), where jumps and trends appear at different times in each filter; also a long-term trend is included. In these light curve a transit pattern with period P=520 time units and a relative depth $\Delta Flux = 0.01$ is included. The transits are masked by the high noise. As can be seen in Fig. 4, all jumps are removed and the resulting output light curve shows some regions with higher noise and some others with lower noise, but this does not affect the real signal. Applying transit detection algorithms (e.g. Box Least Squares - BLS Kovács et al. (2002)), the included transit pattern is also detected.

4. Results

In order to illustrate the algorithm with real light curves, *CDA* is applied to four CoRoT light curves, i.e., CoRoT0102702789, CoRoT0102874481, CoRoT0102741994 and CoRoT0102729260.

4.1. The case of CoRoT0102702789

In Fig. 5 we show the raw red light curve which includes a trend and jumps and the final light curve after applying *CDA* with 5 loops. The light curve of CoRoT012702789 has one huge jump around $JD \sim 2614$ and many other smaller jumps. The RSD_R value of the raw light curve is 5.048 and the final light curve is 0.95. Table 2 shows analytically the values of *RSD* from the total light curves, in these 10 loops of each filter. The green filter has the minimum value and thus *CDA* uses it as a base. The red filter on the other hand has the maximum value and we call it target, but in principal *CDA* defines different filters as base or target in each loop. For this reason in the first four loops the target is the red filter and base the green filter, then target changes to blue and green remains as base etc.; as already mentioned, the red light curve as the most common filter to search for transits.

The example of CoRoT012702789 shows us how *CDA* works and how it removes jumps from a distorted light curve. As far as we can tell from out reconstructed light curve, there are no clear flares or transits in the light curve of CoRoT012702789. The critical question at this point is how *CDA* works if the raw light curve has real events like transits.

Table 2. CoRoT01270289. Table 2 shows how *RSD* is changing in each loop. In the first four loops, red filter is the target and green the base. In loop five this situations has changed. Blue is the target now and green is the base. These values refers to the *RSD* values of the full light curve of each filter.

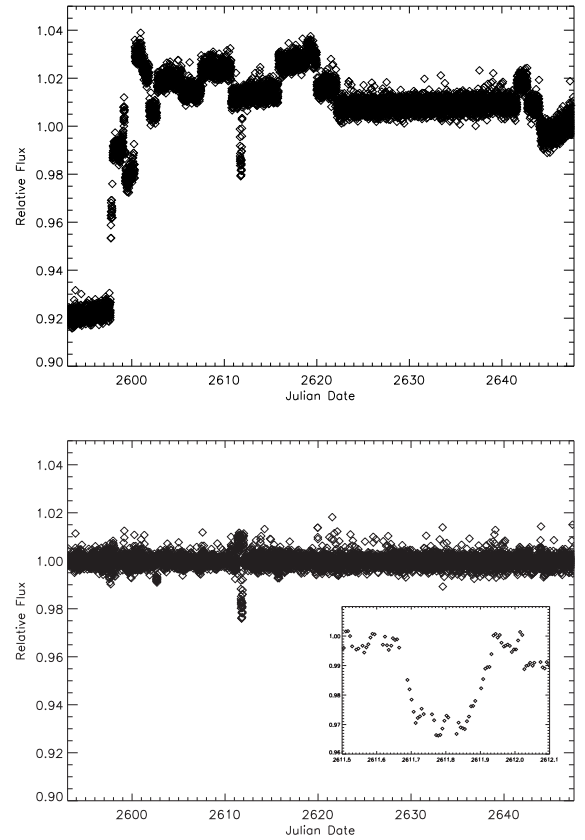| Loop No	RSD_R	RSD_G	RSD_B
♯1	**5.0485**	**0.9497**	1.0658
♯2	**1.8632**	**0.9497**	1.0658
♯3	**1.0665**	**0.9497**	1.0658
♯4	**0.9688**	**0.9497**	1.0658
♯5	0.9688	**0.9497**	0.9868

Fig. 6. CoRoT012874481 - red filter Top: Raw light curve. Bottom: The same light curve after *CDA*. Jumps are removed and a clear transit is appearing. The subframe is a zoom-in plot.

4.2. The case of CoRoT0102874481

An even more extreme case is CoRoT0102874481. The light curve of which is affected by many jumps; the raw (red) light curve of CoRoT0102874481 is shown in Fig. 6. In the raw data it is very difficult to distinguish real from instrumental events. As demonstrated in Fig. 6, *CDA* corrects all the jumps except for a real transit around $JD \sim 2612$. The standard deviation before and after *CDA* is 2203.13 and 336.44 ADUs, respectively. Only a small jump from green and blue filters remains at the end of light curve.

Because this transit is the only transit in the light curve, we cannot determine the period and the nature of the transiting object. Fig. 7 shows that *CDA* does not reduce the depth of the transit, which is ~ 0.036. According to the CoRoT team
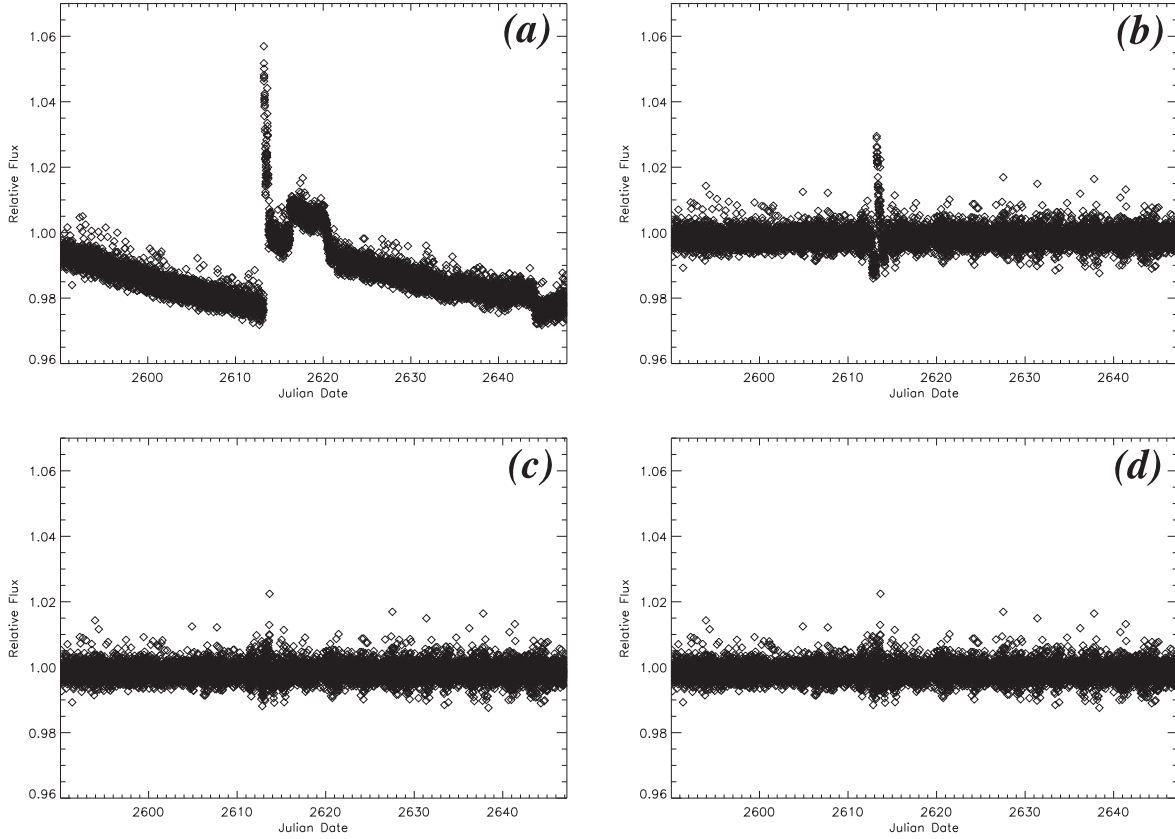
Fig. 5. CoRoT012702789 red light curve and *CDA* results. Raw data - **(a)**, after 1 - **(b)**, 3 - **(c)**, 5 - **(d)** loops respectively. All jumps are removed.

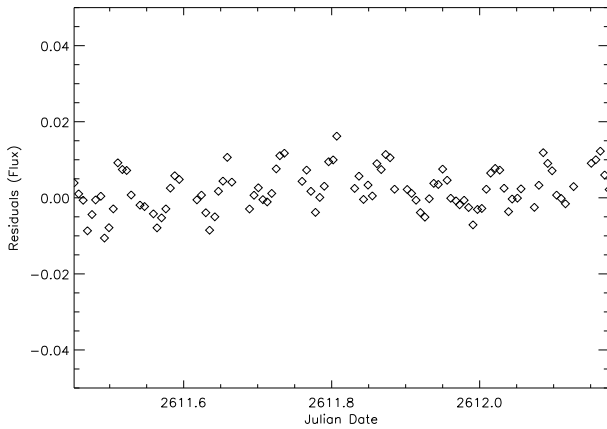**Fig. 7.** CoRoT012874481 residuals before minus after *CDA*. The signal from the real transit is not reduced by the algorithm.

(http://idoc-corot.ias.u-psud.fr), the host star's spectral type is A0IV. Assuming the typical radius and mass of such a star as $R_s = 4.4 R_o$ and $M_s = 2.8 M_o$ and assuming the transiting object to be a true exoplanet, we determine the planet's radius as $R_p = 4.28 R_J$ by using the relation between radius and transit depth (Seager & Mallén-Ornelas 2003)

$$R_p = R_s \cdot \sqrt{\Delta Flux}, \tag{8}$$

where R_s is the radius of the star and R_p is the radius of the planet. From Kepler's 3rd law the semi-major axis of the orbit is $\alpha > 0.78 AU$, because the period is $P > 60$ days.

4.3. The case of CoRoT0102741994

CoRoT0102741994 seems to be a binary system. Our main interest in this example is not to check if *CDA* can remove the jump but to check how the algorithm preserves the eclipses and the flux of the light curve. Fig. 8 shows how the algorithm converts the light curve. The light curve is affected only by a week jump ($\Delta Flux \sim 1.25\%$) around $JD \sim 2615$. The flux depth of the primary and secondary eclipse is 9% and 7%, respectively.

At the top figure is the light curve of the star before the application of *CDA*. The two eclipses are obvious, while the bottom figure shows the light curve after application of *CDA*. Clearly, the jump is removed completely. The depth of the primary and secondary eclipses now are 9.5% and 6.5% respectively. As a general result we can say that *CDA* does not remove the real signal but corrects the jumps.

4.4. The case of CoRoT0102729260

Finally, the case of CoRoT0102729260, is a combination of strong and weak jumps and trends. The raw light curve of CoRoT0102729260 does not show any transits. It is interesting to note that a transit detection algorithm like BLS does not detect any transit event in this light curve (Fig. 10, top panel). However, having applied *CDA* to remove all jumps, we implement again BLS on the final light curve and a possible transit appears (Fig. 9, bottom panel).

This transit is only detectable after applying *CDA*, but not in the raw data. Our analysis of the phased light curve suggests are period of $P = 1.6986$ days. The photometry by the CoRoT team (http://idoc-corot.ias.u-psud.fr) provides some information
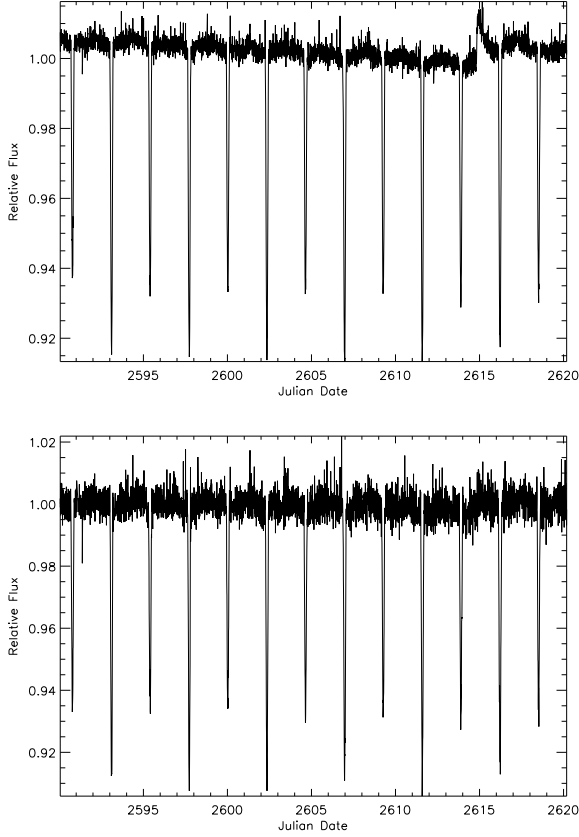
Fig. 8. CoRoT012741994 - red filter Up: Raw data. We just remove all the "bad points". The light curve suffers from one jump around $JD \sim$ 2615 and a trend. Down: The same light curve after *CDA*. The jumps is reduced. *CDA* does not effects the transit depth.

Table 3. Physical Parameters of CoRoT0102729260.

Color Index	0.752
Star Radius R_s	$0.91 R_o$
Period	1.6986 *days*
Planet Radius R_p	$6.27 R_E$
Depth (Flux)	0.004

for the parameter of the host star, which appears to be a main sequence star (G5V) of apparent brightness $m_V = 14.772$ mags. Assuming the spectral type to be correct, we can estimate the radius of the star $R_s \sim 0.91 R_o$. With a transit depth of $\Delta Flux = 0.004$, we deduce a planetary radius of $R_p = 6.27 R_E$ applying Eq. 8. Fig. 11 shows the phase folded light curves. Also Table 3 gives some additional information of the system.

5. Conclusions

We have introduced and presented a method dubbed *CDA* that removes instrumental artefacts from CoRoT data and demonstrated its usefulness in some practical applications. We emphasize that the *CDA* algorithm prepares CoRoT data for any transit detection; it should not be used for transit analysis since it is contingent to remove some real signal. Of course this is not a problem for the detection inasmuch instrumental jumps destroy much more the light curve. From our study of 1030 light curves in the first CoRoT field (IRao01) we found that only very few
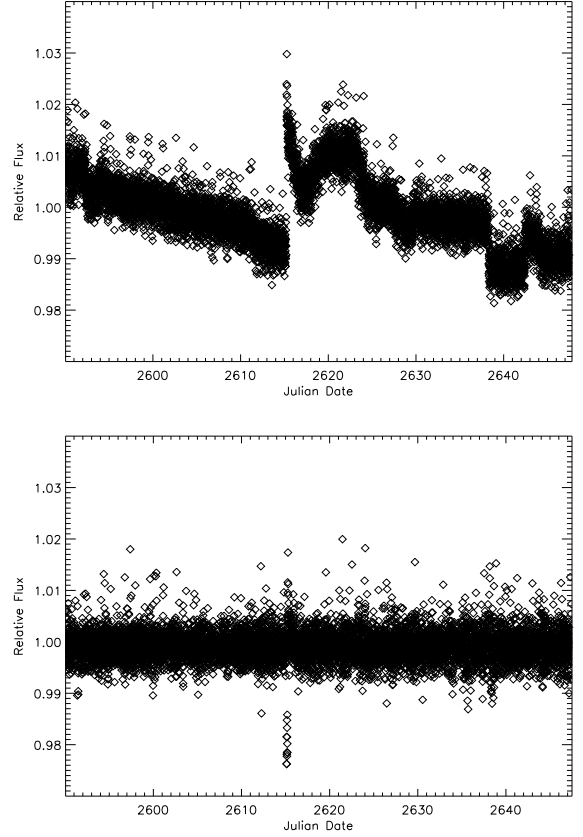
Fig. 9. CoRoT0102729260 - red filter. Up: Raw data before *CDA*. Down: Final light curve after *CDA*. The algorithm succeed to remove all the jumps and trends and improve the light curve enough to detect the "concealed" transit.

light curves have no instrumentally caused features and remain as they are, while the vast majority of light curves are appreciably improved. We present some examples which show how the algorithm affects the light curves. Our main theme is that instrumental jumps substantially affect the CoRoT light curves, making a transit detection in fainter stars impossible.

In order to present how the algorithm affect the full sample, we calculated the Median Absolute Deviation (MAD) before and after appling *CDA*. Fig. 12 shows the differences between the two procedures.

We prove our case with the example of CoRoT0102729260, a possible candidate exoplanet which is detected only after applying *CDA* on the raw data.

References

Aigrain, S., Pont, F., Fressin, F., et al. 2009, åp, 506, 425
Baglin, A., Vauclair, G., & The COROT Team. 2000, Journal of Astrophysics and Astronomy, 21, 319
Kovács, G., Zucker, S., & Mazeh, T. 2002, åp, 391, 369
Mazeh, T., Guterman, P., Aigrain, S., et al. 2009, åp, 506, 431
Pinheiro da Silva, L., Rolland, G., Lapeyrere, V., & Auvergne, M. 2008, MNRAS, 384, 1337
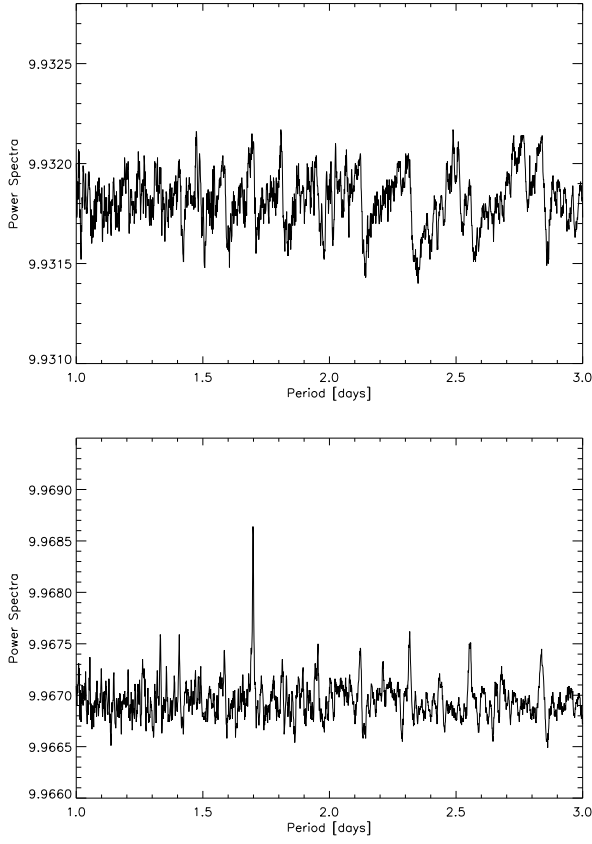Seager, S. & Mallén-Ornelas, G. 2003, ApJ, 585, 1038

Fig. 10. CoRoT0102729260 - red filter. Up: Periodogramm of the raw light curve before *CDA* without any obvious signal. Down: Same plot after *CDA*. A clear periodic signal ($P \sim 1.698$) is detected.
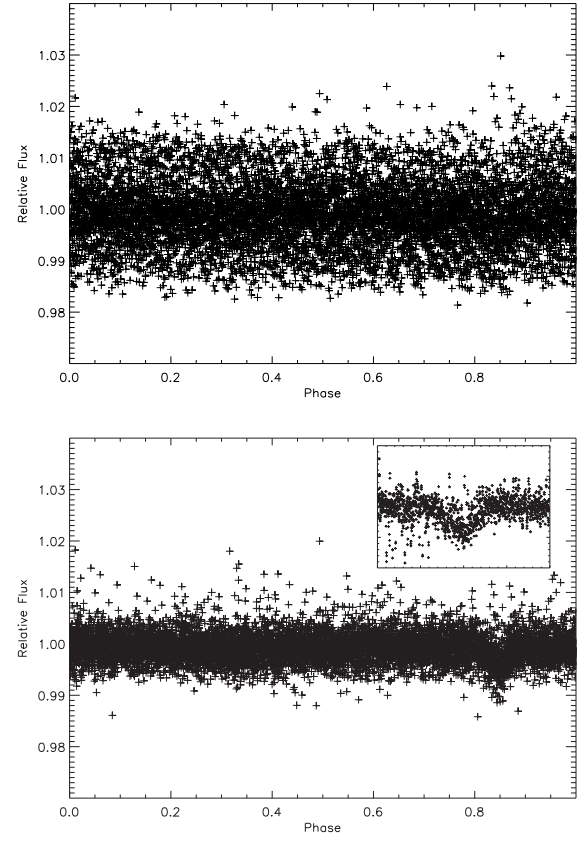

Fig. 11. CoRoT0102729260. Top: A Phase folded light curve before *CDA*. Bottom: A phase folded light curve after *CDA*.
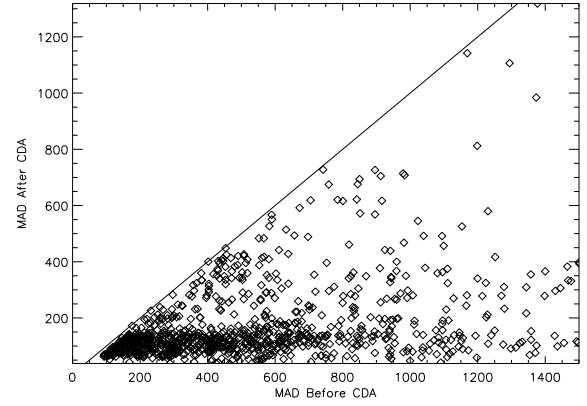

Fig. 12. Median Absolute Deviation (MAD) before and after *CDA* using 1030 lightcurve sample.